DOCUMENT RESUME

ED 244 990 TM 840 313

AUTHOR Takai, Ricky T.
TITLE Test Item Bias in the HS&B Test Battery.
PUB DATE 24 Apr 84
NOTE 13p.; Paper presented at the Annual Meeting of the American Educational Research Association (68th, New Orleans, LA, April 23-27, 1984).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Achievement Tests; Data Analysis; High Schools; Mathematics Achievement; Reading Achievement; Research Problems; *Test Bias; *Test Items; Vocabulary Skills
IDENTIFIERS *High School and Beyond (NCES)

ABSTRACT

The purpose of this papeer is to provide an empirical analysis of four subtests of the High School and Beyond (HS&B) sophomore test battery. This analysis of item bias uses the 1980 test data of the sophomore cohort. Blacks and whites were matched on the total score of four subtests: Vocabulary, Reading, Mathematics--Part I, and Mathematics--Part II. The analysis revealed very little bias. With the exception of two items on the Vocabulary Test and one item on the Mathematics I test, the delta values for the test items on each test were quite low. It should be emphasized that, in this study, the use of the term "bias" refers to those items that behave anomalously in the context of the other items. (BW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED244990

3

TM

TEST ITEM BIAS IN THE HS&B TEST BATTERY

Dr. Ricky T. Takai
Longitudinal Studies Branch
National Center for Education Statistics

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Sylvia Butler
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

TM 840 3/3

Paper presented to the 1984 Annual Meeting of the American Educational Research Association, April 24, 1984.

TEST ITEM BIAS IN THE HS&B TEST BATTERY

With the availability of 1980 and 1982 test data for the HS&B sophomore cohort, researchers are now able to pursue many different types of analyses on student achievement. Undoubtedly, many of these studies will include a comparison of test scores among racial/ethnic groups as a key component of the analyses. In any analyses of test data, it is important to address the issue of item bias. Researchers must be careful to try to separate the "cultural" differences between the various groups from other factors such as high school program and time spent on homework when analyzing HS&B test data. The present paper examines test item bias in the HS&B sophomore test battery.

PURPOSE

The purpose of this paper is to provide an empirical analysis of four subtests of the HS&B sophomore test battery. This analysis of item bias of the HS&B test uses the 1980 test data of the sophomore cohort. It is important to differentiate this type of analysis from what Shepard (1980) calls "logical analysis" of item bias. Logical analysis is defined as "the formal elaboration of what is to be measured and how this trait is expected to relate to other variables". A common method which uses logical analysis is the formation of an expert panel to review the test and pick out those items which do not reflect the intended construct. These "biased" items are then removed from the final version of the test. While the use of logical analysis in detecting item bias is a valuable and important step in constructing an unbiased test, it is not a sufficient method to assess the con-

struct validity of the test. As Cronbach (1971) states, "logical analysis puts forth a counter hypothesis whose pertinence can be verified only empirically". Thus, empirical analysis can sometimes reveal faulty conceptualizations of a trait which may produce different inferences for different groups.

In the past several years, a number of statistical techniques have been developed to perform empirical analyses of item bias (see Ironson and Subkoviak, 1979 for a comparison of the different methods). The common purpose of these different methods is to detect items which are "anomalous." That is, biased items are those which behave differently from the other test items. Following the concept of construct validity, item bias (as defined by empirical analysis) can be defined as an item which does not measure the construct the test was intended to represent.

It should be noted that this definition of item bias avoids the question of bias in the sense of unfairness. As defined here, biased items are viewed as anomalous in a context of other items, not in the unfair way the tests may be used or interpreted. Empirical analysis of item bias may only be interpreted as a way of measuring the construct validity of the test. Thus, bias as "anomalous empirical behavior" of an item is the way in which the term is used in this study.

METHOD

The analysis uses 3064 black and 3035 white students taken from the first follow-up sample of the HS&B sophomore cohort. In order to adjust for

differences in overall achievement levels between blacks and whites, the two groups are matched on the total score of the four subtests used in this analysis: Vocabulary, Reading, Mathematics - Part I, and Mathematics - Part II. The method of matching involved comparing the frequency distributions of test scores for blacks and whites. Random samples of students were selected at each score level for the group with the larger frequency until the frequency distributions in both groups are about equal. The reason for this procedure is to allow for the detection of items which function differently for persons of the same achievement levels in different groups. By attempting to control for differences in achievement, real group differences in achievement can be distinguished from those created by irrelevant difficulty or differential validity in what the item is assumed to represent. While it is recognized that a limitation of this matching procedure is that it is based on a criterion internal to the tests in question, the author is in agreement with researchers of recent studies (Angoff and Ford, 1973; Ironson and Subkoviak, 1979) who decided it was worthwhile to try to control for differences (by matching on an internal criterion) in achievement in order to separate the "signal" (real differences) from the "noise" (item bias).

As mentioned before, four subtests were used in this analysis. These tests were chosen because they tend to be used by most researchers to measure verbal (Vocabulary and Reading) and quantitative (Mathematics - Part I and II) achievement levels of HS&B students. In addition, it has been argued that traditional items are more closely associated with the cultural experiences of the white middle class. Therefore, items on the traditional vocabulary, reading, and mathematics tests were considered as potentially

more biased.

Each of the subtests is designed to measure a different cognitive skill. The vocabulary test (21 items, 7 minutes) uses a synonym format. The reading test (19 items, 15 minutes) is based on short passages (100-200 words) with several related questions concerning a variety of reading skills (analysis and interpretation), but focusing on reading comprehension. The Mathematics Part I test (28 items, 16 minutes) consists of quantitative comparisons in which the student indicates which of the two quantities is greater, equal, or is impossible to tell. The Mathematics Part II test was similar in format to the Part I test but the level of difficulty of questions is higher.

The method used to detect item bias is the method described in Angoff and Ford (1973) known as the transformed difficulty procedure. The method consists of computing the difficulty or p-value (proportion getting the item correct) for each item for each group. The normal deviates (z) are obtained corresponding to the (1-p)th percentile of the distribution (i.e., z is the value having proportion (1-p) of the normal distribution below it), and then to delta values in order to eliminate negative z-values. The formula for delta is:

$$\text{delta} = 4z + 13.$$

For both blacks and whites, mean and item delta values are calculated using the 1980 test battery data. Delta values are plotted (black vs. white) for each test. The distance (d_i) of the delta points (one point for each item)

from the major axis of the ellipse created by the delta points is taken as the measure of that item's bias. The choice of this particular method of detecting bias is made for several reasons. First, the transformed difficulty procedure is easy to interpret. The concept of item difficulty (proportion answering the item correctly) is easy to grasp. Also the ability to visually display those items which appear to be biased by delta plots simplifies the analysis of the results. Second, the transformed difficulty procedure is easy to implement. While the newer and more sophisticated techniques such as the item characteristics curve method may be more theoretically appealing, the requirements of long subtests, large sample sizes, and special computer software (LOGIST, ANCILLES) make it difficult to implement. The transformed difficulty procedure however does not have the above requirements and delta values can be calculated by hand.

RESULTS

Means and standard deviations of item deltas for blacks and whites on the four HS&B subtests are given in Table 1. Overall, there is very little difference in the mean delta values between the black and white samples on the four tests. Average delta scores are slightly higher for blacks on the vocabulary, mathematics I, and mathematics II tests; whites have a higher score on the reading test. As with previous studies, the matching procedure appears to have brought the two groups closer to each other in terms of overall levels of item difficulty. Differences between mean delta values for blacks and whites are less than two-tenths of one standard deviation (of the black sample) on the four subtests.

These small differences can be contrasted with the formula test scores for the entire samples of black and white sophomores:

	RACE			
HS&B TEST	BLACK		WHITE	
	X	SD	X	SD
Vocabulary	4.98	4.70	9.96	5.03
Reading	4.35	3.85	7.78	4.72
Mathematics I	5.17	6.08	11.86	7.55
Mathematics II	1.27	2.12	2.99	2.80

As shown above, the mean test scores for the total black sample are much lower than for the white sample. Means are about one standard deviation (using the black sample as the base) lower for blacks than whites. Because the achievement levels of blacks and whites are different on these four subtests, the attempt to control for overall differences in achievement levels by matching appears warrranted.

In Table 2, we begin to examine the degree to which the items have the same rank order of difficulty in the two groups by presenting correlation coefficients between the item deltas. As mentioned by Angoff and Ford, an item delta correlation is also an inverse expression of the item x group interaction. The correlation between the item deltas of the two groups reveal that the vocabulary test appears to have the greatest discrepancy in terms of rank ordering of the items (.97). Surprisingly, the other component of verbal achievement - reading, has the highest correlation (.99)

of the four groups. The two parts of the mathematics test have intermediate correlations of .98. The finding that the correlation between deltas is higher for the mathematics II than vocabulary test is also surprising since we might expect that differences in rank ordering of item difficulty might be greater for tests that might be affected by differences in curriculum content. Since the mathematics II test was designed to contain more difficult items than the mathematics I test and verbal learning is not quite as confined to the classromm as is mathematics learning, the lower correlation on the vocabulary test is unexpected.

It should be noted that while the range of correlation coefficients is quite small (.97-.99), the standard deviation of the distances from the delta points to the major axis of the ellipse increases by proportionately large amounts as the correlation departs from unity by what appears to be small amounts. Using the forumla given in Angoff and Ford to calculate the standard deviation of the distances of the points:

$$s_D = 1 - r_{xy},$$

we find that the value of s_D for the vocabulary test is .17 almost twice as large as the s_D for the reading test (.10).

The final section of the analysis turns to the comparison of item delta values for each subtest. In Tables 3A through 3D, delta values are presented for each item for each group. The distance (d_i) of the delta point from the major axis of the ellipse is also given for each item. Accompanying

these tables are the delta plots (Figures 1 through 4) for the subtests. The line drawn through the points represents the major axis of the ellipse created by the delta points.

In Table 3A we see that items nos. 4 and 11 have the largest deviations from the major axis. Item no. 4 appears to be biased in favor of whites, while item no. 11 is biased in favor of blacks. For item no. 4, the student was asked to link a synonym to the word "beat"; for item no. 11, a synonym for the word "compliment". Two other items which have moderately large values of d_i are nos. 8 and 17. Again, one item (no. 8) is biased against blacks and the other (no. 17) against whites. With the exception of these four items, the magnitude of the distances of the points from the axis is quite small (see Figure 1).

Table 3B shows a few items which can be considered to be moderately biased (nos. 1 and 20). In both items, blacks are more likely to have difficulty with these items than whites. However in general there does not appear to be strongly biased items in the reading test. Figure 2A provides a good illustration of the lack of bias in this subtest as the delta points are closely positioned to the axis.

One item which appears to be biased in favor of the black sample on the Mathematics I test is item no. 28, the last item. The item deals with comparing a common fraction versus a decimal fraction. It is not clear whether this bias can be attributed to cultural differences or to some other factor such as the fact that this item does not contain any accompanying text. Items nos. 5 and 14 appear to be moderately biased in favor of

whites, while item no. 17 biased in favor of blacks.

The Mathematics II test appears to have no items that are distinctively biased in favor of one group. As Figure 4 shows, the delta points lie close to the axis of the ellipse. Items nos. 5, 6 and 8 (in favor of whites) have only moderate d_i values.

SUMMARY

The item bias analysis of the the 1980 test data of the HS&B sophomore cohort reveals very little bias in the four subtests. With the exception of two items on the vocabulary test and one item on the mathematics I test, the d_i values for the test items on each test are quite low. Again, it should be emphasized that the use of the term "bias" refers to those items that behave anomalously in the context of the other items. The notion of whether the item is "bad" does not apply to our analyses. The empirical analyses of test item bias as it is presented here only attempts to identify those items which do not appear to be measuring the underlying construct of the test.

As with most research studies, the ubiquitous and vague final phrase "more work needs to be done" applies to this study. However there are two specific studies which could directly follow this work. First, item bias analysis should be done for multiple groups of students. In particular, Hispanic and Asian students are two groups where cultural differences might

be reflected in test performance. Second, the large sample size of HS&B will allow for the more sophisticated item bias methods to be attempted. Following the work done by Ironson and Subkoviak on the NLS-72 test battery, a similar type of analysis could be done for the 1980 sophomore cohort test battery.

REFERENCES

Angoff, W.H. and Ford, S.F. "Item-race interaction on a test of scholastic aptitude." Journal of Educational Measurement, 1973, 10, 95-105.

Cronbach, L.J. "Test Validation." In R.L. Thorndike (Ed.) Educational Measurement (2nd ed.) Washington, D.C.: American Council on Education, 1971.

Ironson, G.H. and Subkoviak, M.J. "A comparison of several methods of assessing item bias." Journal of Educational Measurement, 1979, 16, 209-225.

Shepard, L.A. "Definition of Bias." Paper presented at the Johns Hopkins National Symposium on Educational Research. Washington, D.C., November, 1980.